DATA ANALYTICS И СИСТЕМАТА ORANGE

Даниела Орозова
Бургаски свободен университет

DATA ANALYTICS AND ORANGE SYSTEM

Daniela Orozova
Burgas Free University

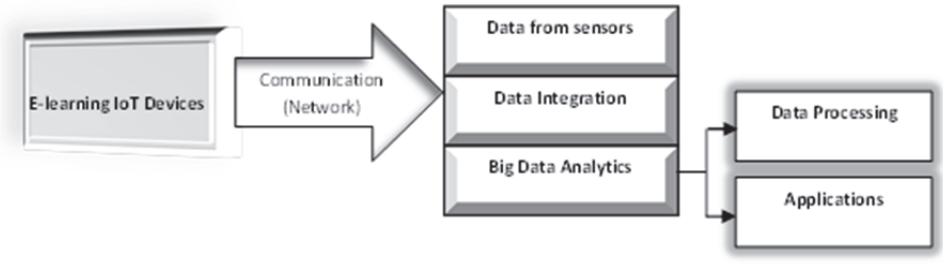
Abstract: *Основната цел на тази статия е да покаже въздействието на Data Analytics в областта на образователно пространство. Представят се базови възможности на инструментите за анализи и визуализация на данни на системата Orange. Приложени са редица обучаващи примери за изследователски анализ на данните.*

Key words: *Data Analytics, Data Science, Virtual Education Space, E-learning.*

1. Науката за данните

Днес с навлизането на Интернет на нещата в нашия живот нараства нуждата от анализ на данните за извличане на полезна информация. Например за по-добро разбиране и насочване към клиентите, компаниите могат да допълват своите бази от данни с нови данни от социални мрежи, браузъри, сензорни и др., за да получат по-пълна картина на своите клиенти. Банките и телекомуникационните компании могат по-добре да предвидят загубата на клиенти; търговците могат да определят кои продукти ще се продадат заедно; автомобилните застрахователни компании могат да проучат колко добре клиентите им шофират и т.н.. Основната цел е да се създават прогнозни модели, на базата на които да се взимат управленчески решения.

В съвременния живот, данните, генерирани от IoT сензорите, се предават чрез мрежата и интегрираните данни се анализират [6], функционален изглед на процеса е показан на фигура 1.



Фигура: 1. IoT и Data Analytics

Термините Big Data, Data Analytics и Data Mining описват както самите данни, така и технологиите за събиране, обработка, управление на данните и методите за анализ. Data Mining е процеса на търсене на скрити данни и закономерности, предва-

рително неизвестни, нетривиални и практически полезни, необходими за взимане на решения в различни сфери на човешките дейности. Тук акцентът е не само в извличане на нови факти, но и генериране на хипотези, които могат да бъдат проверявани. Традиционните инструменти на анализа се основават на математическата статистика – регресия, корелация, клъстеризация, анализ на времеви редове, дървета на решенията и др., а също и техники на изкуствения интелект като: машинно обучение, невронни мрежи, генетични алгоритми, размити логики и др.

Big Data Analytics се явява развитие на концепцията Data Mining. Също така е и развитие на решаваните задачи, сфери на приложение, източници на данни, методи и технологии на обработка. От появата на концепцията Data Mining до настъпване на ерата на Big Data, се изменя обема на анализираните данни, появяват се високопроизводителни системи, нови технологии, в това число Map/Reduce и нейните многочислени програмни реализации. Появява се науката за данните – Data science.

Data science съчетава множество подходи и техники, свързани с анализ на данни от областта на статистиката, откриване на знания, машинно обучение, изкуствен интелект, програмиране, комуникация др. Науката за данните включва и процесите по изчистване и интеграция на данните, избор и трансформация на данни, извличане на знания, техния анализ, оценяване и представяне. Може да се каже, че *Data science* е „сплав“ от различни дисциплини, технологии и средства за анализ на данните.

Независимо дали целта е да се открият интересни взаимовръзки, да се категоризират обекти в групи, да се оптимизира планирането на ресурси или да се определят тарифи за таксуване, основното разбиране на техниките за анализ на данните, може да помогне за извличане на полезни знания и коректно решаване на задачите.

При задачата за класификация е зададено крайно множество от обекти, за които е известно към кои класове принадлежат, а класовата принадлежност на останалите обекти е неизвестна. Трябва да се построи алгоритъм, който класифицира произволен обект като укаже стойността на целевия атрибут. Когато възможните стойности на целевия атрибут са само две, то имаме „бинарен” класификационен проблем, в другия случай – „многокласов”. Най-общо алгоритъмът работи като създава серия от случайни правила. При задачата за предвиждане, предварително данните са разделени на две групи с взаимно изключващи се елементи – тестови и тренировъчни групи данни. На базата на първото множество се изгражда модел на данните, като се генерират правила и се избират тези, които отговарят най-добре на данните. Процесът се повтаря определен брой пъти, докато се намери правило, което да удовлетворява (почти 100%) тренировъчните данни. След това правилата се проверяват чрез тестовите данни и се оценява модела.

Регресионният анализ дава отговор на въпроса какви са причините. Той показва взаимните отношения между величините, които могат да бъдат интерпретирани като причинно-следствени. Това е статистически анализ, предназначен да дава количествен израз на ефектите на дадена група променливи X1, X2, ..., Xp, които условно се наричат „независими“ върху друга променлива Y, която се нарича „зависима“. Основната идея е търсене на естествена функционална връзка от вида: y = f (x1 , x2 , , xp). Уравнението се нарича уравнение на регресия.

Асоциативен анализ – изучава честотата на съвместно появяване на факти. Този анализ е свързан с откриване на „асоциативни правила”, задаващи условия за стойностите на атрибутите, които се явяват често заедно в дадено множество от данни. Асоциативните правила имат вида: X => Y. Така правилото се състои от две части: X е условната (предшестваща) част, а Y е логическото следствие (резултантна част) [2].

При клъстерния анализ – целта е n на брой обекта да се групират в k на брой групи, наречени клъстери, като се използват p на брой признаци (променливи). Така клъстерът се формира от подобни обекти, независимо от техните класове. Целта е разкриване на евентуално скрита групировка на обектите. Едно важно деление на клъстеризационните процедури е в зависимост от това, дали се задава предварително броя на клъстерите. Голямото разнообразие на процедурите се поражда от използваните правила за създаване на клъстерите [3]. По-известни са методите „на най-близкия съсед", „на най-отдалечения съсед", „на центроидите" и др.

Анализ на шумове – в данните могат да се съдържат обекти, които не поддържат основното поведение или модел на данните. Те се наричат отклонения (Outliers) и се определят като шумове. Понякога, обаче, тези данни може да са по-интересни от останалите случаи. Отклоненията са разликите между измерените стойности и съответните очаквания на базата на предишни или нормативни стойности. Задача за анализ на данните е при откриване на множество от отклонения да се създаде описание на характеристиките на отклоненията, да се обясни причината за това, да предложи действие за довеждане на стойностите обратно към техните очаквания и др.

2. Средства на системата Orange за визуализация и анализ на данни

Софтуерът **Orange** [15]. е разработка на факултета по биоинформатика в университета на Любляна. Той е с отворен код, базиран на езика Python. За процеса на инсталиране на продукта е нужно да се изтегли подходящата версия от уеб сайта: https://orange.biolab.si/download/. Например: Classic installer (Default), Orange3-3.4.5-Python34-win32.exe.

Orange е базирана на компоненти, визуална среда за програмиране. Компонентите на Orange (widgets) предоставят широк спектър от възможности: от елементарна визуализация на данни, предварителна обработка и валидация до оценяване на алгоритми за обучение и изграждане на модели за прогнозиране [5].

Следва разглеждане на редица примери за визуализиране и изследователски анализ на данните, чрез които се представят основните възможности на системата Orange. В началото на работата, създаваме нов работен процес, след което се зарежда работния плот (canvas) и набора с инструменти. Всяко извличане на данни започва със зареждане на данните чрез инструмента „*File*", които са предварително изчистени и във формат за работа в „Orange". Нека да изберем инсталирания файл с данни „*iris.tab*" от галерията на системата. Данните могат да бъдат въведени от Excel (.xlsx), от текстов файл с раздели (.txt), файл с данни разделени със запетая (.csv) или URL адреси.

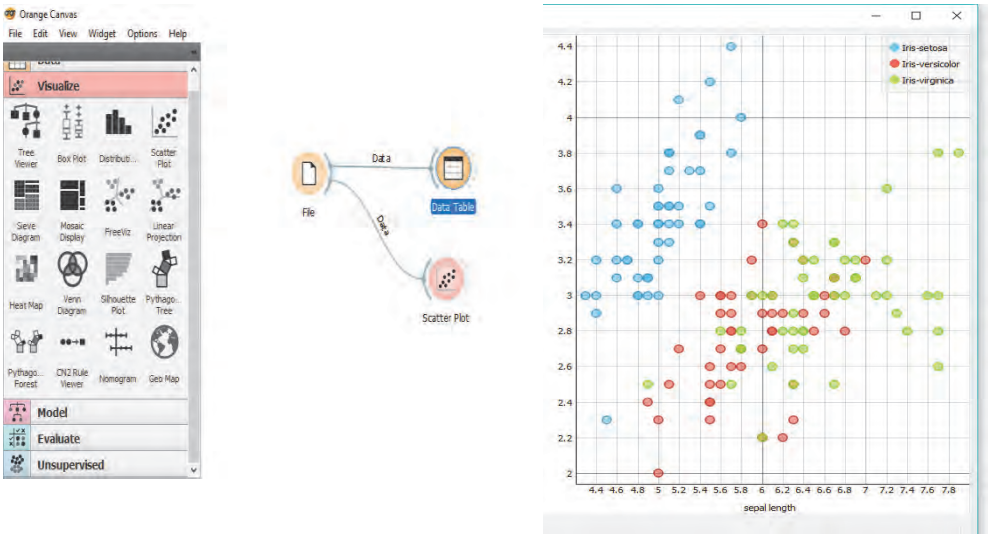
Във файла „iris.tab" са въведени данни, относно размерите на венчелистчетата и чашелистчетата на ирисите. След като са прочетени данните от файла, с два клика върху изхода на „*File*" избираме да създадем връзка с „*Data Table*", така изпращаме данните за инспекция. На фигура 2 е изобразена връзката за комуникация между двата инструмента „*File*" и „*Data Table*".

Фиг. 2. Инструменти „ File" и „Data Table"

Инструментите за визуализация включват: „*scatter plot*", „*box plot*", „*histogram*", „*distributions*", „*silhouette plot*", „*tree viewer*", „*pythagorean tree*" и редица други. Предлагат се и допълнителни опции за визуализации на мрежи, географски карти, изображения и др. За да представим нагледно приложението на инструментите, разглеждаме последователност от примери:

- **Scatter Plot**

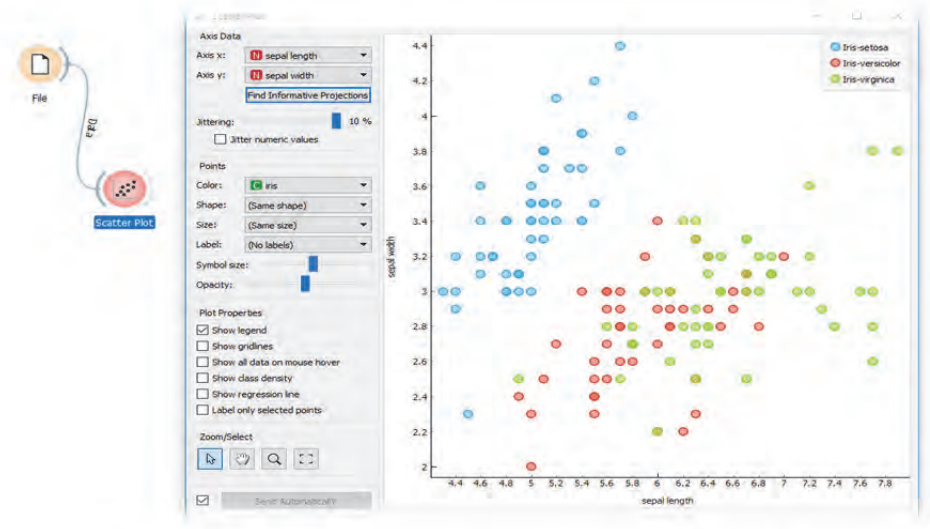
Инструментът *Scatter Plot* осигурява двуизмерна визуализация, както за непрекъснати, така и за дискретни признаци. Данните се показват като набор от точки, всеки, от които има ***x-axis*** стойност, определяща позицията по хоризонталната ос и ***y-axis*** стойност на атрибута, определящ позицията по вертикалната ос. На фигура 3 са представени стъпките за визуализация на данни с инструмента *Scatter Plot* на „Orange".



Фиг. 3. Визуализация на данни с инструмента Scatter Plot на „Orange".

Различните свойства на графиката, като: цвят, размер и форма на точките, заглавията на осите, максималния размер на точките и трептенията, могат да се коригират от диалогов прозорец на инструмента, показан на фигура 4. Отново използваме файла, относно данните за ирисите „iris.tab".

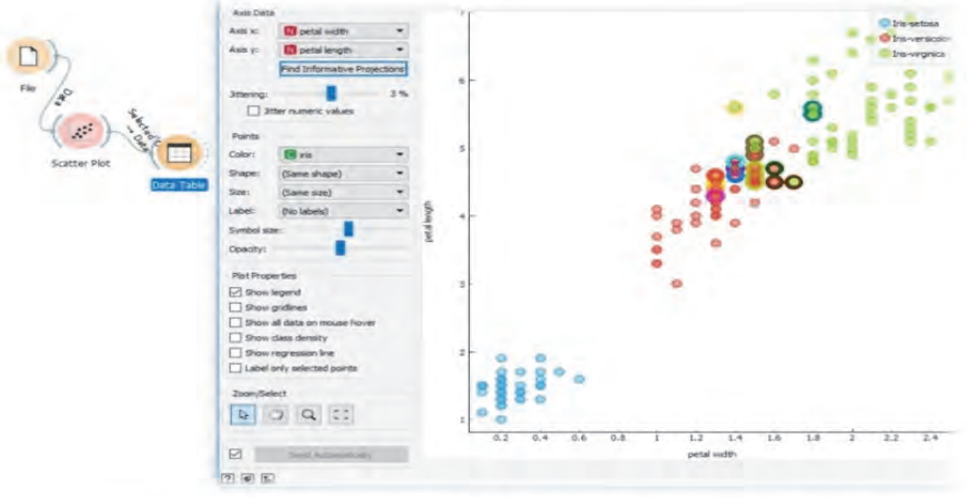
Ако даден набор от данни има много атрибути е трудно ръчно да се сканират всички двойки, за да се намерят интересни или полезни връзки. Затова в „Orange" може да използва визуализация на данните с опцията *Find Informative Projections* (Start Evaluation). Функцията ще върне списък с двойки атрибути по средна оценка на точността на класификация. Целта на оптимизацията е да се намерят проекции на *scatter plot*, където случаите са добре разделени.



Фиг. 4. Различни свойства на графиката с инструмента „Scatter Plot“

Scatter Plot, както и другите инструменти на „Orange“, поддържа възможности за приближаване и изключване на част от графиката, както и ръчен избор на селекция за данни. Тези функции са достъпни в долния ляв ъгъл на прозореца на инструмента. Инструментът по подразбиране е *Select*, който избира данни в област. *Pan* позволява да се премества плота около избрания прозорец. С мащабиране може да се увеличава мащаба, докато *Reset zoom* нулира визуализацията до оптимален размер.

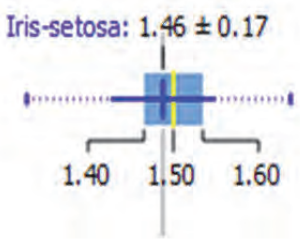
Scatter Plot може да се комбинира с всеки друг инструмент. На фигура 5 е показана извадка от данни в *Scatter Plot*. Данните са селектирани със задържане на бутона Shift и кликване с ляв бутон на мишката, от *Data Table*.



Фиг. 5. Селекция на данни в Scatter Plot, които са представени в Data Table

- **Box Plot**

Инструментът *Box Plot* показва разпределението на стойностите на атрибутите. Добра практика е да се проверяват всички нови данни с този инструмент, за да се открият бързо несъответствия, като дублирани стойности, големи различия в стойностите или подобни данни. В резултат на анализ на **атрибут с непрекъснати стойности**, можем да видим изображение от типа, показан на фигура 6:



Фигура 6. Резултат от работа на инструмент Box Plot в „Orange”.

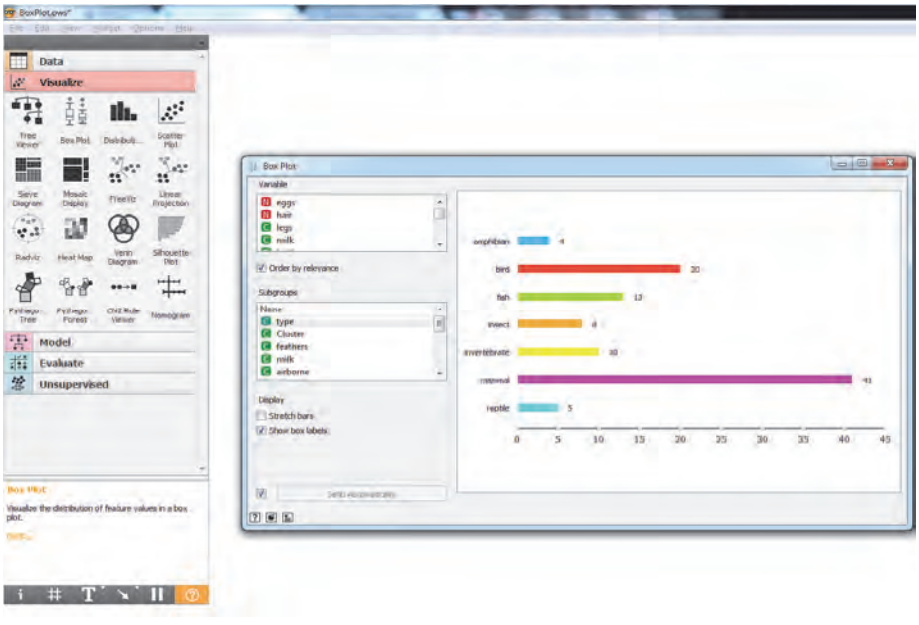
Това изображение представя следните статистически стойности:
- Средата (тъмносиня) вертикална линия представя модата.
- Светлосиния правоъгълник представя стандартното отклонение от модата (стандартно отклонение на средната стойност).
- Медианата е представена с жълтата вертикална линия.
- Непрекъснатата хоризонтална синя линия показва разликата между областта на първата част (от 25%) и останалата част (от 75%) от данните.
- Тънката пунктирана линия представя целия диапазон от стойности (от най-ниската до най-високата стойност в набора от данни за избрания параметър).

На фигура 7 са визуализирани данните от файла „iris.tab”, като се анализират стойностите на дължините на чашелистчетата на ирисите.

Фиг. 7. Визуализация на резултата от анализа на данни с инструмент Box Plot.

За **атрибути с дискретни стойности**, лентите представляват броя на случаите с всяка отделна стойност на атрибута. Нека сега да заредим файла „Zoo.tab“ от галерията на Orange, в този случай *Plot*-инструментът, (фигура 8) показва броя на различните видове животни в набора от данни за зоологическа градина. Има 41 бозайници, 13 риби, 20 птици и т.н.

Box Plot е полезен за намиране на статистическите зависимости на конкретен набор от данни. Например набор от потребителски данни, дефинирани в друг инструмент (като Scatter plot) или обекти, включени към някой клъстер или възел за класификация.



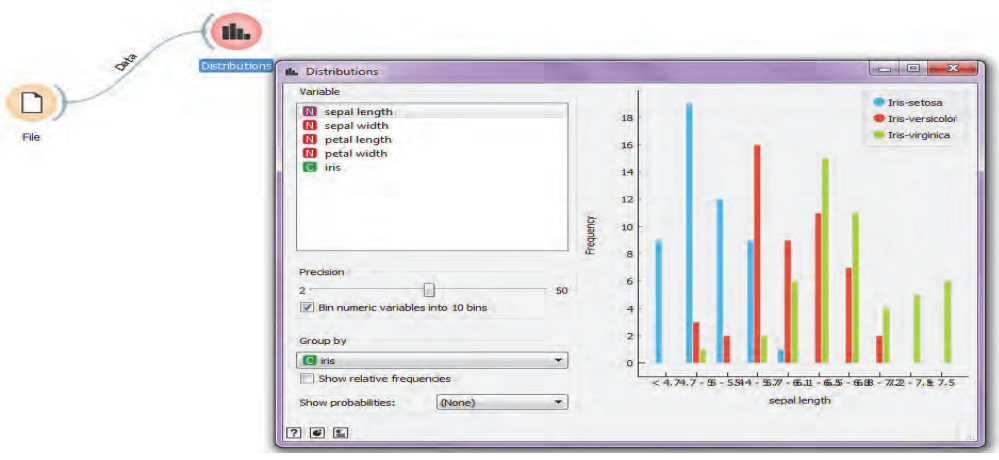
Фиг. 8. Визуализация на дискретни данни с инструмента Box Plot в „Orange”.

- **Distributions**

Инструментът *Distributions* показва разпределението на стойностите на атрибути с дискретни или непрекъснати стойности.

За **дискретни атрибути** графиката, отразява колко пъти всяка стойност на атрибута се появява в данните. На фигура 9 е показана графика, като са използвани данните от “iris.tab”, сравнявайки стойностите на дължините на чашелистчетата за трите класа ириси.

Инструментът *Distributions* показва разпределенията на стойностите за указани променливи. Ако се маркират променливи с непрекъснати стойности, *Bin* инструментът ще дискретизира променливите, като ги зададе на интервали. Броят на интервалите се определя със зададена точност. Друга възможност е да зададем гладкост на кривите на разпределение на непрекъснатите променливи. От инструмента *Distributions* може да бъде поискано да показва раздели със стойности само за случаи от определен клас, използвайки *Group by*. Също могат да бъдат показани и изчислените вероятности чрез *Show probabilities*.



Фиг. 9. Визуализация на данни с инструмент Distributions.

При **непрекъснати атрибути** стойностите се показват като функционална графика. Класовите вероятности за непрекъснати атрибути се получават с оценка на плътността на гаусовото ядро, а появата на кривата се определя от лентата с инструменти - *Precision* (*smooth* or *precise*). В следващия пример, показан на фигура 10, отново се използва набора от данни за ирисите (Iris.tab), като е създадена графика на данни за атрибут с непрекъснати стойности, визуализирани с инструмент *Distributions* в системата „Orange".

Фиг. 10. Визуализация на данни с инструмент Distributions.

- **Pythagorean Tree**

Използва се за визуализиране и изследване на дървовидни модели. След изграждане на модел, чрез инструмента Tree, резултатът може да се визуализира с инструмент *Pythagorean Tree*, използвайки равнинни фрактали, които изобразяват общите йерархии на дървовидната структура. На следващата фигура 11 е представена връзката между *File*, *Tree* и *Pythagorean Tree* инструментите и визуализация на данните в системата „Orange", използвайки файла с данни *„iris.tab"*.

В диалоговия прозорец се задават параметрите за визуализацията:

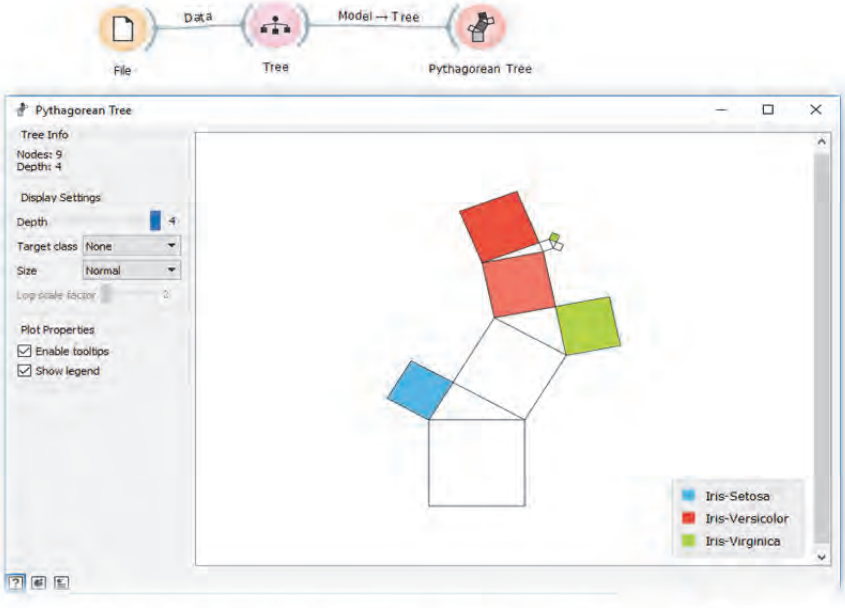
- *Depth* задава дълбочината на показваните дървета;

- *Target class* указва целевия клас за класификационното дърво, като интензитетът на цвета на възлите съответства на вероятността за целевия клас;

- *Size* задава възможност за изчисляване на размера на квадрата, представляващ възела. Ако е избран *Size-Normal* се запазват размерите на възлите, съответстващи на размера на подгрупата данни от възела. Ако е избран *Size-Square root* то се задава квадратен корен от съответната трансформация на размера на възела. При *Size-Logarithmic* трансформацията на възела е логаритмична.

- *Plot properties*: опцията *Enable tooltips* дава възможност за показване на информация относно всеки възел;

- *Show legend* – при активиране се извежда легенда.

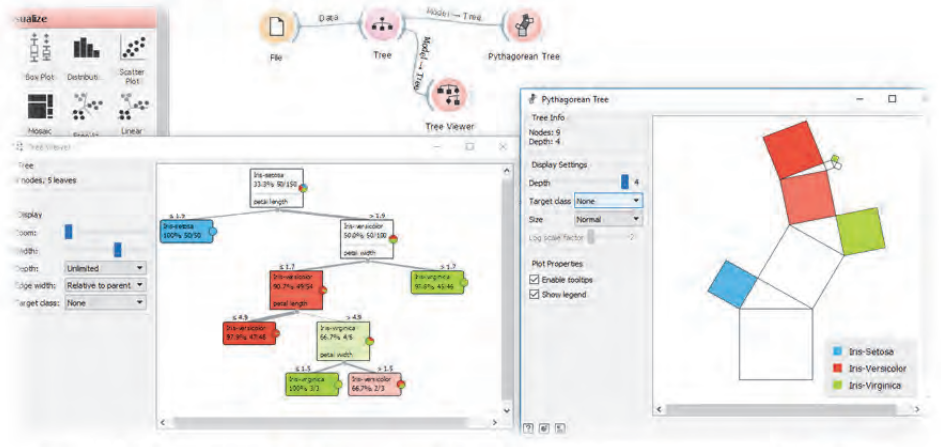


Фиг. 11. Визуализация на данни с инструмент Pythagorean Tree.

- **Tree Viewer**

Tree Viewer е друг инструмент за визуализация на дървовиден модел.

Tree Viewer и ***Pythagorean Tree*** са инструменти, които визуализират дърво, но визуализацията на Питагор изисква по-малко пространство и е по-компактна, дори и за малък набор от данни (фигура 12).



Фиг. 12. Визуализация на данните от файла Iris.tab.с инструментите Tree Viewer и Pythagorean Tree

- **Sieve Diagram**

Този инструмент построява диаграма с правоъгълници, като площта на всеки правоъгълник е пропорционална на очакваната честота за избраната величина, докато наблюдаваната честота е илюстрирана чрез броя на квадратите в него. Разликата между наблюдаваната и очакваната честота се проявява в честотата на защриховане, използвайки цвят за индикация. Ако отклонението има положителна стойност се използва син цвят, ако е отрицателно – червен цвят.

Класификационен анализ със средствата на „Orange“

Класификацията е процес на разпределение на дадено множество от обекти в съответствие с приетите в системата правила. В резултат се получават подмножества, обединяващи част от обектите на изходното множество на основание един или няколко признака. Такива подмножества се наричат класификационни групи.

Йерархичната класификация е система за класификация, която се използва когато класификационните признаци за определена номенклатура са йерархично зависими и подчинени. При тази класификация се получават съподчинени класификационни групи. Получените подмножества не се пресичат. Последователността на разделяне е следната: отначало класификационното множество се дели по избран признак на големи подгрупи. Всяко новосформирано множество се класифицира в зависимост от стойностите на следващ признак и по подобен начин множеството се дели последователно до необходимата степен на детайлизация. При прилагането на този метод, връзките между членовете на дадена класификация образуват йерархия между елементите, която графично може да се представи като *дърво*.

В следващия пример нека из**ползваме база от данни за пасажерите на „Титаник“**. Исторически данни са достъпни на адрес: https://en.wikipedia.org/wiki/RMS_Titanic. Таблици с данни във формат *.csv са достъпни на адрес: https://www.kaggle.eom/c/titanic/data (*test.csv; train.csv).* За конкретен анализ на тези данни е необходимо те да се „почистят”, подредят и подготвят, като се запазят само потенциално полезните. В редица случаи е необходимо данните да се преформатират и декомпозират. В практиката се използват различни инструменти за тези цели, например:

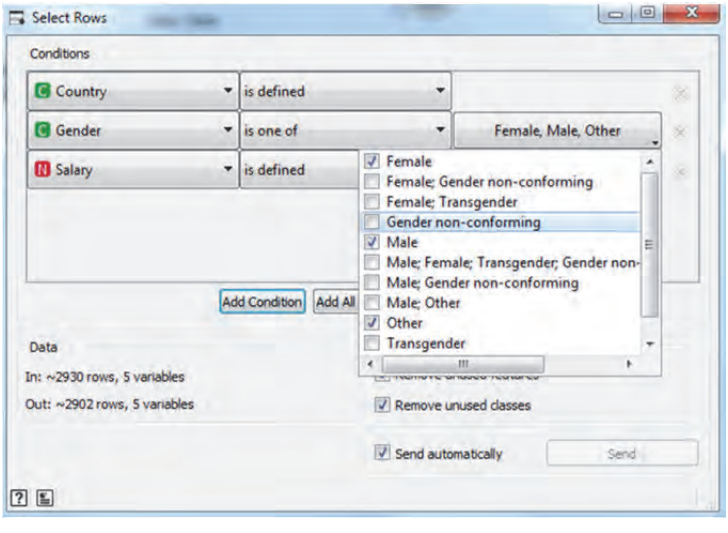
- Excel – мощен инструмент за работа с таблични данни, в който се поддържат редица функции за обработка и анализ на данните.
- Релационни СУБД – в тази категория попадат системи, в които могат да се заредят таблични данни и да се разпределят в таблици с релации между тях. Този подход дава голяма гъвкавост при филтрирането и редуцирането на липсващи данни и изпълнение на сложни заявки върху данните.
- Скриптови езици – най-използвани за тази цел са езиците R, Python, Perl. Те са мощни инструменти при работата с големи обеми от данни и позволяват провеждане на мащабни анализи.

В някой случаи, когато се обработва огромно количество данни се налага преобразование, за да се редуцира представеното множество данни, без това да промени резултата от работата с тях. Редукцията на данните може да се реализира по различни начини като:

- редукция на размерностите – премахване на някои несъществени за анализа размерности;
- компресиране на данните – с цел намаляване на физическия им размер;
- заместване на данните с по-малки по размер алтернативни данни, чрез параметризация;
- създаване на концептуални йерархии, чрез представяне на данните в групи (клъстери) на различни йерархични нива.

- **Select Rows**

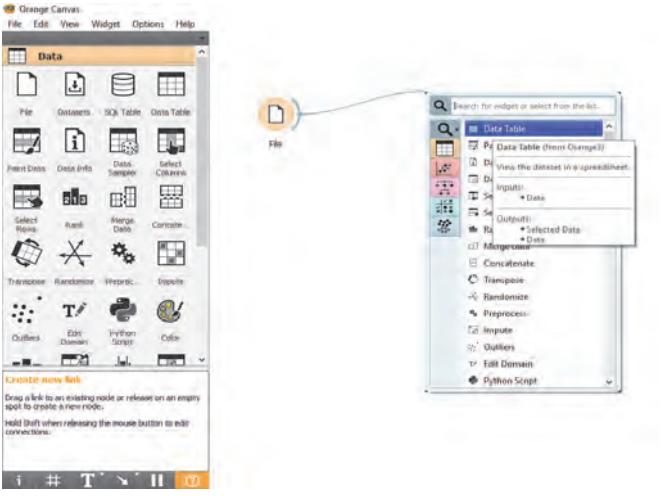
Чрез системата Orange можем да филтрираме данните, например с помощта на инструмента *Select Rows*, като едновременно с това идентифицираме колоните, по които бихме искали да получим извадка, върху която ще правим анализ. Можем да изберем само конкретни стойности на атрибут, които искаме да участват в анализа. Например за атрибута пол са въведени различни стойности, ние указваме на системата за този атрибут да приема единствено възможни стойности – Male, Female и Other (фигура 13).



Фиг. 13. Работа с инструмента Select Rows.

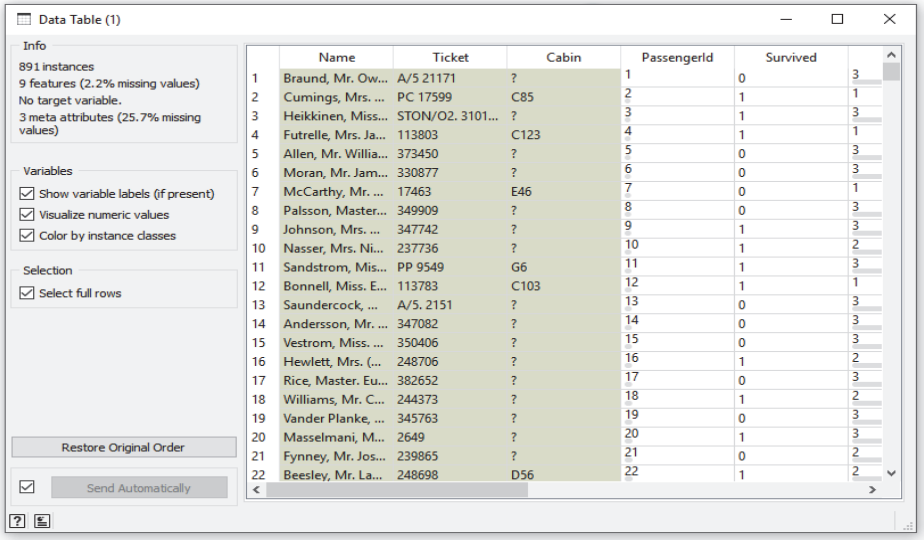
За решаване на задачата, свързана с йерархично клъстеризиране, извършваме следната последователност от стъпки:

1. Избираме файла с данни, в случая ***train.csv***
2. Зареждаме избрания файл с данни с помощта на инструмента ***File***.



Фиг. 14. Избор на инструмент в системата Orange.

3. В прозореца на инструмента File се визуализира информация за заредения набор от данни: размер, брой и типове данни.
4. Преглед на данните с инструмента ***Data Table.*** Инструментът *Data Table* получава един или повече набори от данни в своя вход и ги представя като таблица. Данни могат да бъдат сортирани по стойности на указани атрибути.



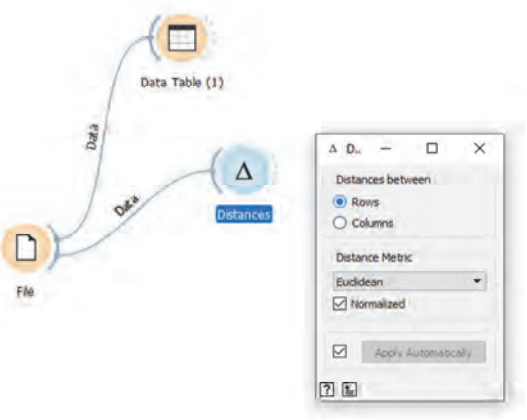
Фиг. 15. Преглед на данните от файл с инструмента Data Table.

5. Създаване на връзка: ***File*** -> ***Distances***.
6. Прилагаме йерархично клъстеризиране с инструмент ***Hierarchical Clustering***.
7. Изготвяне на отчет (***Report***).

- **Distances**

Инструментът ***Distances*** изчислява разстоянията (между редовете или колоните) в даден набор от данни. От инструмента ***Distances*** трябва да изберем метрика за разстоянието *Distance Metric*. Задаваме *Euclidean* – познатото евклидово разстоянието между две точки.

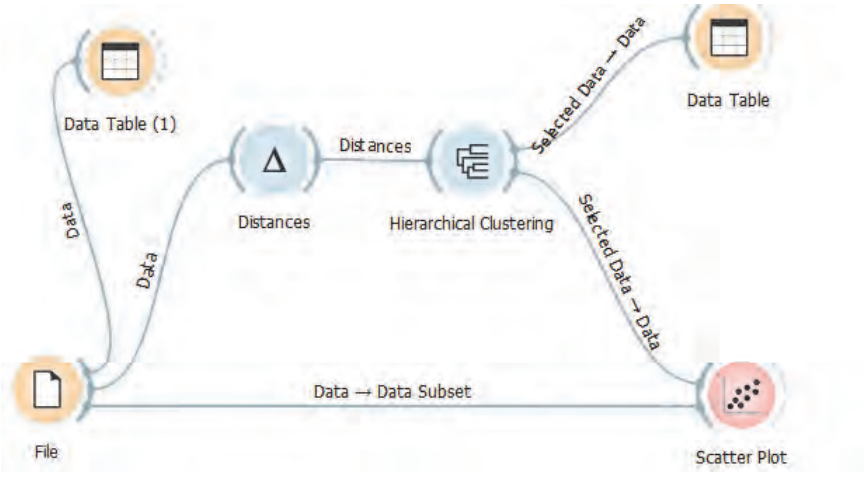
Този инструмент *Distances* трябва да бъде свързан с друг инструмент, който изисква измерване на разстояния (степен на близост) на обекти. В примера използваме *Hierarchical Clustering* за групиране на обектите.



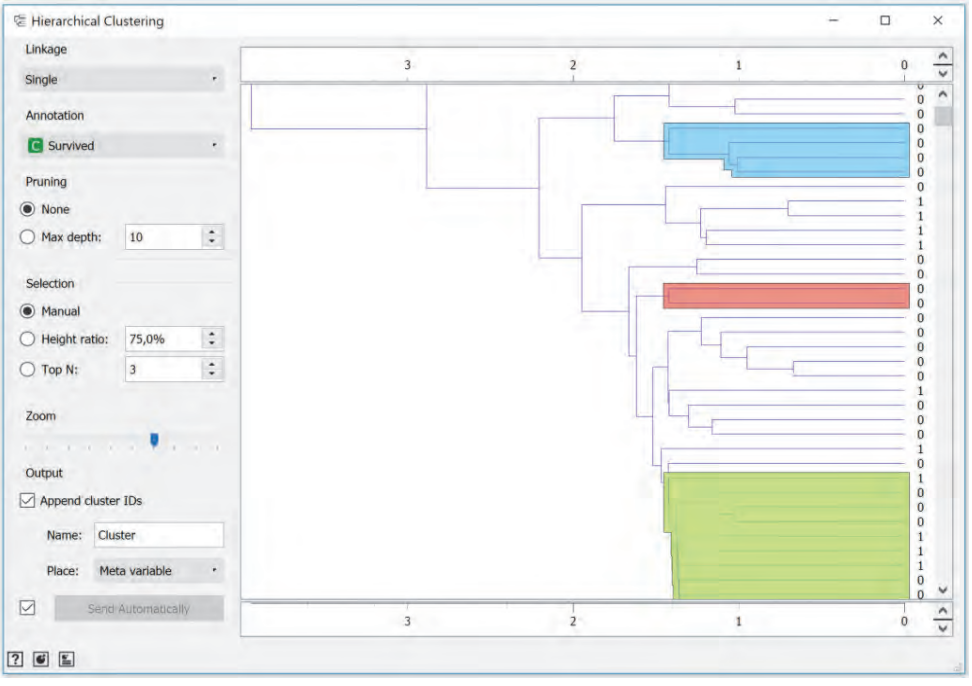
Фиг. 16. Инструмент Distances.

- **Hierarchical Clustering**

Инструментът ***Hierarchical Clustering*** извършва йерархичното групиране на произволни типове обекти чрез изчислените разстояния *Distances* и показва съответната дендрограма. *Дендрограмата* е граф-дърво, в което всеки възел отразява една стъпка от процеса на клъстеризиране. Тя носи и допълнителна информация за разстоянието между двата клъстера.



Фиг. 17. Процес на йерархично клъстеризиране и визуализиране



Фиг. 18. Дендрограма – резултат от работата на инструмента Hierarchical Clustering

Инструментът поддържа четири начина за измерване на разстоянията между клъстерите:

- Single linkage – разстоянието между най-близките елементи на двата клъстера;
- Average linkage – средното разстояние между елементите на два клъстера;
- Weighted linkage – претеглено свързване, използвайки метода WPGMA.
- Complete linkage разстоянието между най-отдалечените елементи на клъстерите

Етикетите на възлите в дендограмата могат да бъдат избирани в полето Annotation. Огромни дендрограми могат да се изрежат в полето *Pruning*, като се избере максималната дълбочина на дендрограмата. Това засяга само визуализацията, а не действителното клъстеризиране.

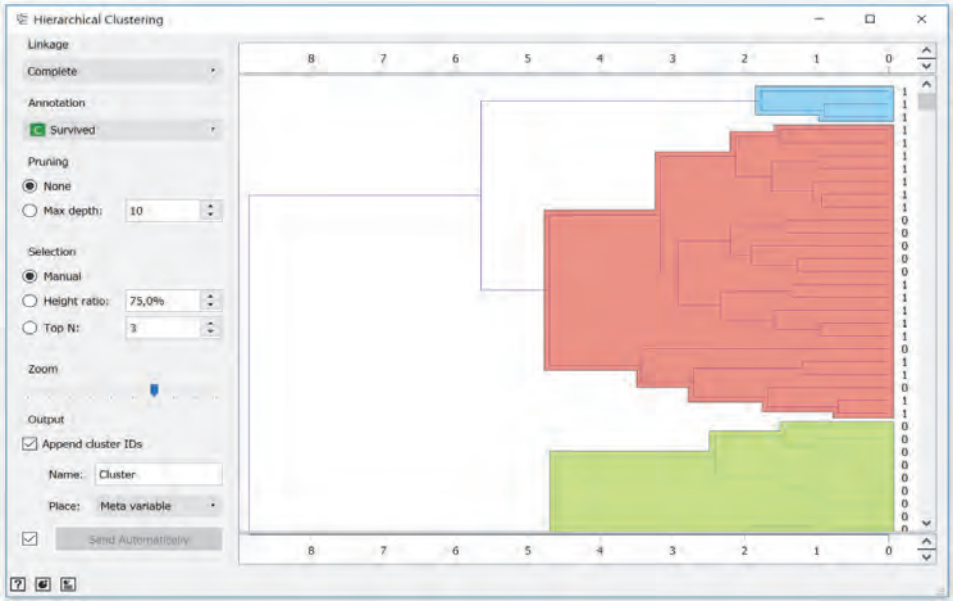
Инструментът предлага три различни метода за избор:

• **Manual (ръчен)** – чрез кликване в рамките на дендрограмата ще се избере клъстер. Няколко клъстера могат да бъдат избирани чрез задържане на Ctrl, като всеки избран клъстер е показан в различен цвят и се разглежда като отделен клъстер в изхода.

• **Height ratio** – задава коефициент на пропорция на височината**.**

• **Top N** – избира се броя на възлите.

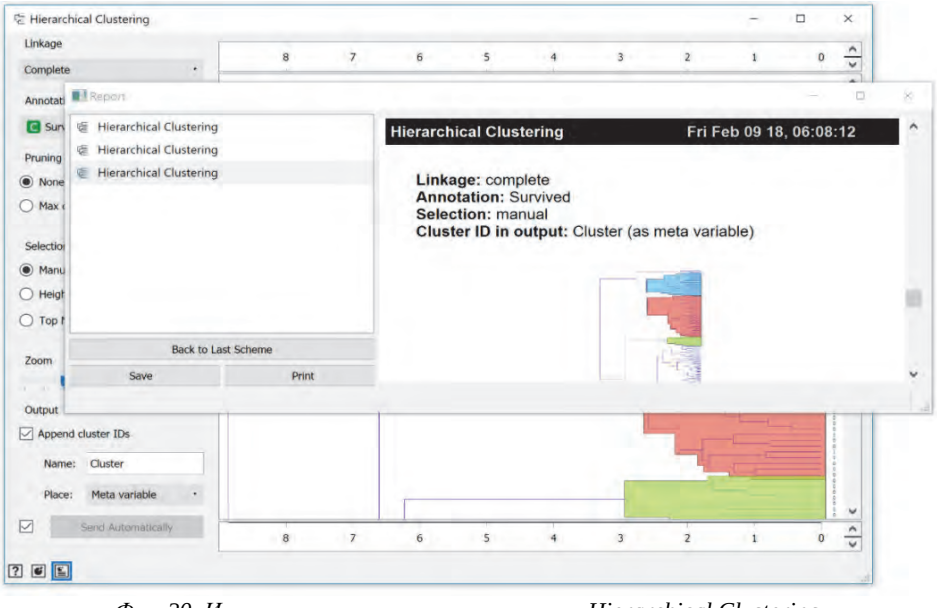
Може да се приложи мащабиране *Zoom*, за да се увеличи или намали мащаба.



Фиг. 19. Настройки на инструмента ***Hierarchical Clustering***

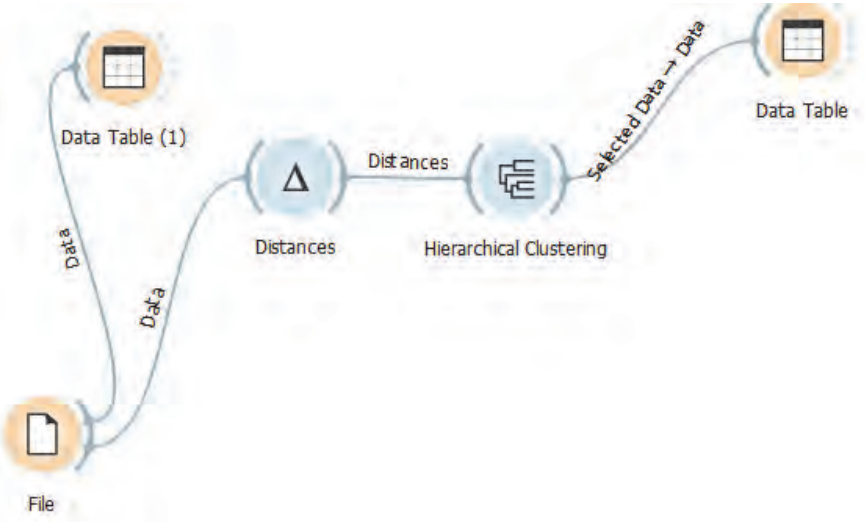
Данните могат автоматично да бъдат изведени при всяка промяна *Auto send* is *on* (когато е включена опцията автоматичното изпращане) или, чрез натискане на *Send Data*. Натискайки този бутон, ще се създаде изображение, което може да бъде запазено.

Последната стъпка е изготвяне на отчет – ***Report*** с инструмента Hierarchical Clustering.



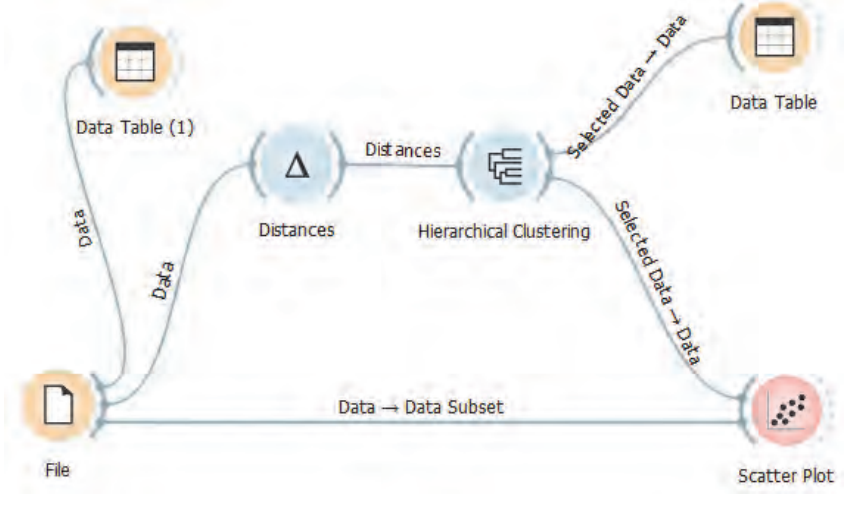
Фиг. 20. Изготвяне на отчет с инструмента Hierarchical Clustering

Ако изберем *Append cluster IDs* – следва добавяне на идентификационни номера на клъстерите и можем да видим допълнителна колона в таблицата с данни, наречена *Cluster*. За целта трябва да добавим още един инструмент *Data Table* след клъстеризирането. Това е начин да се провери как йерархичното клъстеризиране групира отделните случаи.



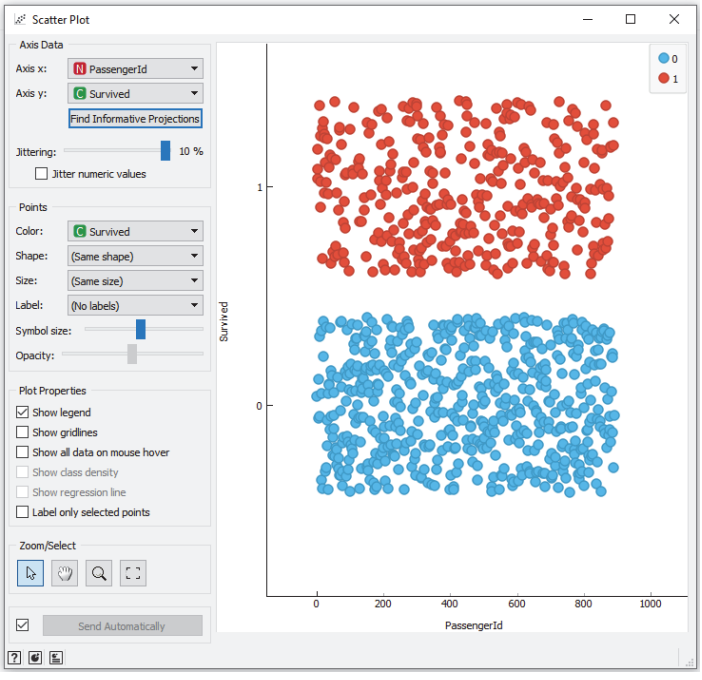
Фиг. 21. Работен поток за йерархична клъстеризация.

Нека сега към процеса добавим и инструмент *Scatter Plot* за визуализация на данните за избрани случаи от йерархичното клъстеризиране (фигура 22).



Фиг. 22. Работен поток за Hierarchical Clustering с инструмента Scatter Plot.

Инструментът *Scatter Plot* осигурява двумерна визуализация. На изображението (фигура 23) е показан набора от данни за атрибут *Survived* за пасажерите на „Титаник“.



Фигура 23. Визуализация на набора от данни за атрибута Survived.

Клъстерен анализ със средствата на „Orange“

Процесът на клъстеризация е предназначен за разбиване на съвкупността от обекти на еднородни групи – клъстери. Колкото повече обектите вътре в клъстера са по-подобни един на друг и се отличават от обектите в другите клъстери, толкова по-точна е клъстеризацията.

Клъстерите обикновено се изобразяват в n-мерно евклидово пространство и се описват чрез съответни геометрични характеристики – координати на центъра (центроида), радиус и т.н. За оценка на сходството на обектите, попадащи в даден клъстер, се използват метрики, които обикновено се основават на евклидовото разстояние между точките или други характеристики.

Методите за решаване на задачата за клъстеризация са две основни групи:

- **Йерархични –** чрез рекурсивно обединяване или разделяне на обектите в клъстери се създава дървовидна структура на клъстеризацията, изобразена като дендрограма. Методите са подходящи за прилагане върху ограничено множество от неголям брой обекти.

- **Нейерархични –** основаващи се на итеративни техники за групиране на изходната съвкупност от данни. Най-популярният метод от тази група е метода k-средните (k-means), известен още като „бърз клъстерен анализ”.

При този метод на К-средните величини (K-Means Cluster Analysis) се отчита разстоянието на всяка единица до центровете на отделните клъстери, като най-

близкото разстояние определя принадлежността на единицата към съответния клъстер. Методът изисква предварително да се определи броят на клъстерите. Центровете на тези клъстери могат да бъдат известни или да се оценят от данните. Освен това центровете могат да останат постоянни или да се актуализират в процеса на анализа.

Задача: да се анализират данни, чрез програмата Orange Data Mining, относно закъснения на полети и да се направи модел, като за целта се използват данните от следния линк: https://data.world/data-society/airlines-delay.

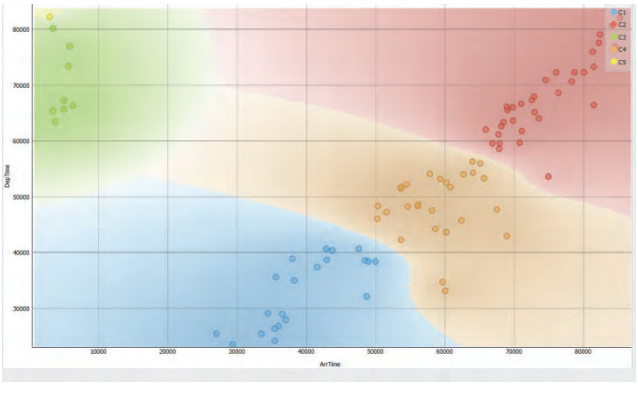
Данните във файла са относно закъснения на полети. При този анализ ще търсим какво общо има между закъсняващите полети, с цел изясняване на причините за закъсненията. Разглеждайки зададения голям обем данни, при първоначален анализ можем да решим да изключим данни, които могат да не са необходими за анализа, като например месец, година и ден от месеца относно полета.

- **Нейерархична клъстеризация k-means**

За решението нека построим модела на k-means, избирайки последователно различен брой клъстери. Установяваме, че след 5 клъстера, няма значително отличаващ се от останалите клъстер. Поради, което се спираме на брой клъстери – 5.

Чрез *Scatter plot* инструмента построяваме възможните проекции и избираме първата, която е с най-голяма вероятност, в случая атрибутите ***ArrTime / DeptTime***.

Получава се следната диаграмата, показана на фигура 24.



Фиг. 24. Визуализация на пет клъстера по атрибутите ArrTime и DeptTime.

- **Йерархична клъстеризация**

Йерархичният клъстерен анализ се провежда на два етапа. Резултатът на първия етап е броя на клъстерите, на които следва да разделим данните. На втория етап извършваме клъстеризацията като използваме броя на клъстерите, които сме определили на първи етап.

Различните методи за свързване на единиците могат да доведат до различни резултати при формирането на клъстери.

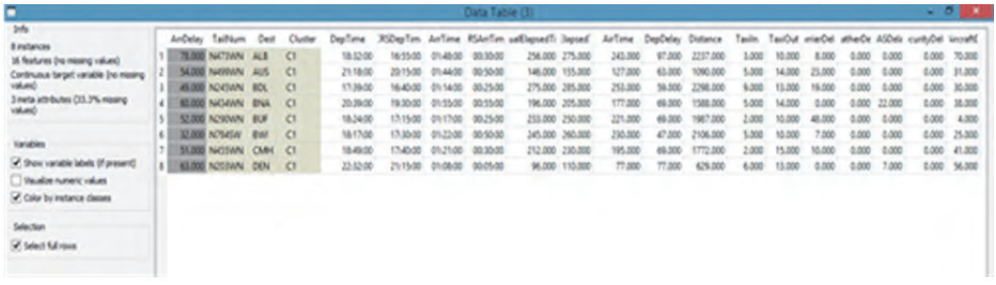
- метод на междугрупова свързаност (Between-groups linkage);
- метод на вътрешногрупова свързаност (Within-groups linkage);
- метод на най-близкия съсед (Nearest neighbor);
- метод на най-далечния съсед (Furthest neighbor);
- центроиден метод (Centroid clustering);

- медиален метод (Median clustering);
- метод на Вард (Ward's method).

Основен инструмент при йерархичната клъстеризация е *дендрограмата*, която представлява графично изображение, показващо процеса на формиране на клъстерите чрез присъединяване на случаите към тях.

Когато изберем даден клъстер от дендрограмата, данните за клъстера се прехвърлят към свързаната таблица с данни (*Data Table*). Получената таблицата може да се отвори и в нея се визуализират данните за полетите, дестинация, причината за закъснение и време, както и номера на клъстера, в който попадат полетите. Извадка от примерна таблица, показваща подробна информация за избран клъстер е дадена на фигура 25.

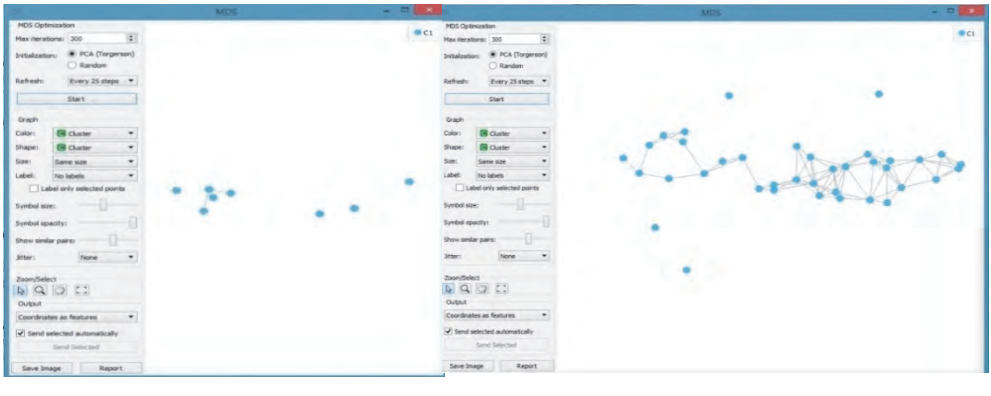
Избирайки опцията *Save Data*, можем да запишем избрани данни от даден клъстер под формата на нова таблица за следващи изследвания.



Фиг. 25. Извадка от таблица с данни, относно избран клъстер.

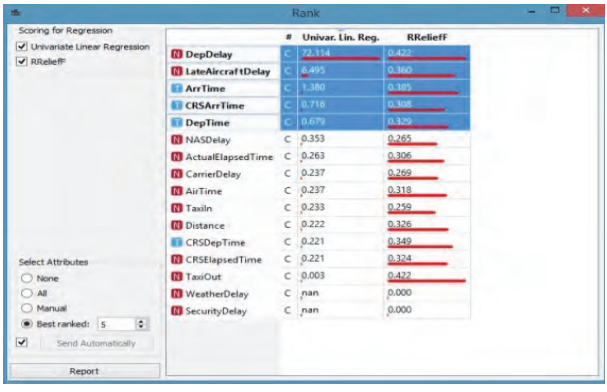
- **MDS проекция**

В работния процес за инструмента *Hierarchical Clustering* можем да прикачим инструмент *MDS – двумерна проекция на* данни, която прави многомерно мащабиране на данните, на базата на матрица с разстояния (фигура 26).



Фиг. 26. MDS графика на два от получените клъстери.

Rank – класифицира и филтрира данните според тяхното значение. Най-важните характеристики са показани на фигура 27.



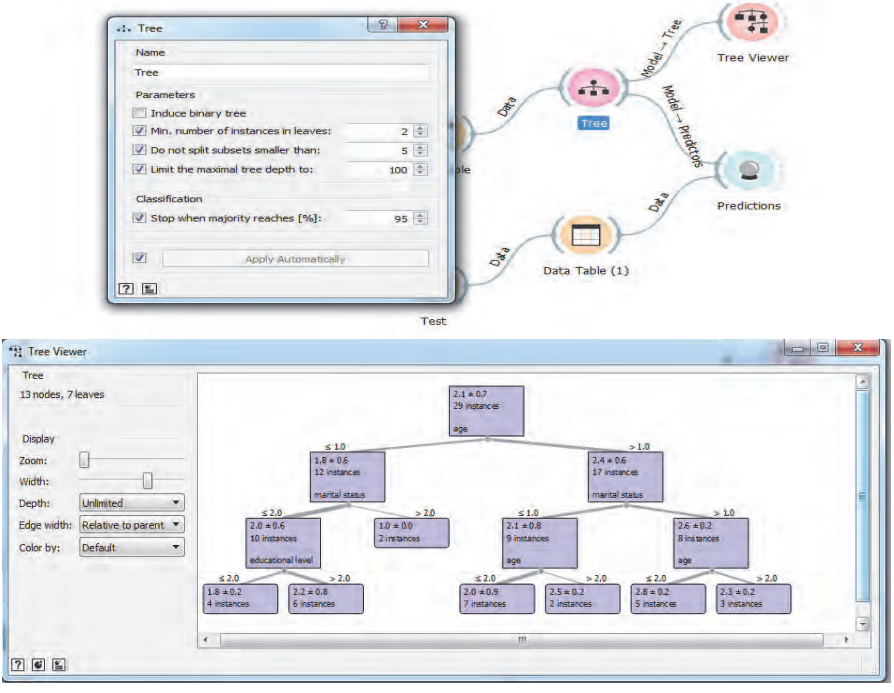
Фиг. 27. Класифициране на данните от клъстерите

- **Venn Diagram**

Инструментът *Venn Diagram* визуализира данните, които се припокриват чрез представяне на множествата от данни с кръгове в различни цветове.

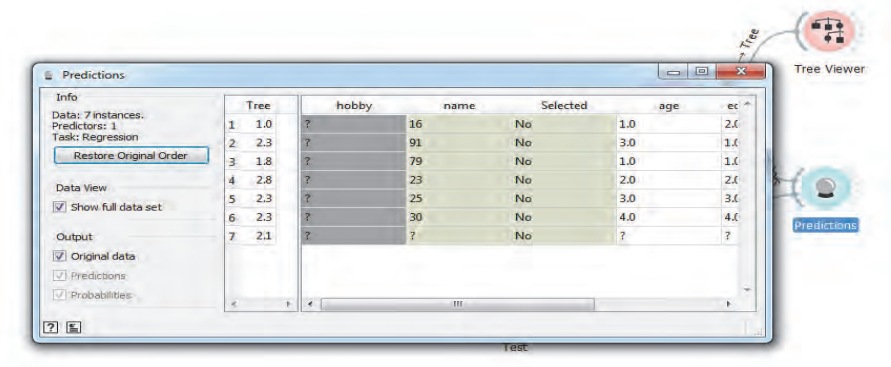
Задача за предвиждане със средствата на „Orange “

Нека използваме файла с данни **Data.xml** (с данни за потребители – имена, възраст, хоби, ниво на образование и материален статус) и да създадем модел с инструмента *Tree*. След това подготвяме таблица **Test.xml** със същата структура, но колоната *хоби* не е зададена (?) и ще бъде предвиждана от създадения модел.



Фиг. 28. Създаване дървовидна структура и визуализиране с инструмент Tree Viewer.

Първо създаваме дървовиден модел чрез инструмента *Tree* и визуализираме дървовидната структура чрез инструмента *Tree Viewer*. Потребителят може да избере възел и по този начин да изведе данните, свързани с възела за анализ. Процесът е показан на фигура 28.



Фиг. 29. Структура на файла Test.xml.

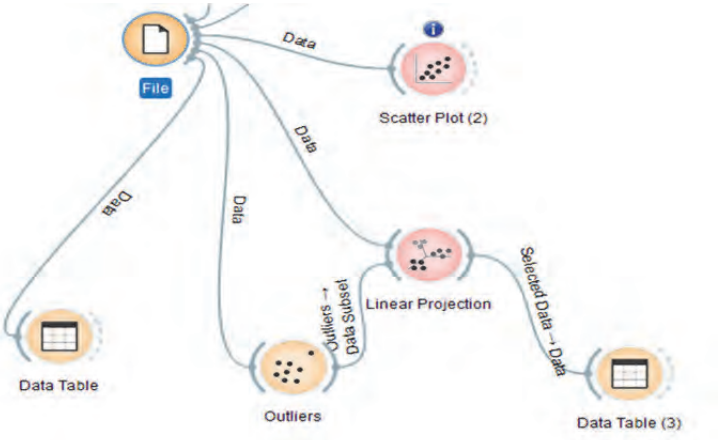
От менюто *Evaluate* избираме инструмента *Predictions*, който извършва прогнозите за данните от файла *Test.xml* и определя стойност на незададената колона – *хоби*, предвиждана от създадения модел. Структура на файла *Test.xml* е показана на фигура 29.

Графичният потребителски интерфейс на системата *Orange* дава възможност на потребителите да се съсредоточат върху изследователски анализ на данните, вместо върху кодиране на алгоритмите. Системата предлага и компоненти за машинно обучение и възможности за разширяване на функционалността за извличане на данни от външни източници, извличане на текст, мрежов анализ и т.н.

3. Извличане и анализ на данните в среда за електронно обучение

Науката за данните в полза на обучението е нов поглед, който съчетава множество подходи за обработка на данни, получавани от различни университети среди и системи. Особен интерес представляват данните за отпадащите студенти от онлайн курсовете за обучение [7]. Тук на базата на данните, събирани за студентите, участващи в различни форми на електронно обучение, търсим връзка между статистически методи, машинно обучение, откриване на модели на поведение и анализи на данните.

Например: избираме файла с данни за студентите и го зареждаме с помощта на инструмента *File*. В прозореца на инструмента се визуализира информация за заредения набор от данни: размер, брой и типове данни. Преглед на данните се извършва с инструмента *Data Table*, а тяхното графично визуализиране с инструмента *Scatter Plot*.



Фиг. 30. Примерен работен процес чрез системата Orange.

Тук бихме могли и да филтрираме данните с помощта на инструмента *Select Rows* и да идентифицираме колоните, върху която ще правим анализ. В случая първоначално разглеждаме дали има стойности, които се различават драстично от останалите чрез инструмента *Outliers*.

Ако приложим класификационен анализ ще се идентифицират подмножества от обекти с общи характеристики. За целите на този анализ включваме инструмента *Distance* за измерване на разстоянието между отделните точки и избираме метрика *Distance Metric – Euclidean* (евклидово разстояние). След прилагане на йерархична клъстеризация с инструмента *Hierarchical Clustering* ще получим групи от студенти с общи характеристики. Накрая изготвяме отчет (*Report*).

Всеки преподавател, който създава електронен курс би искал да направи курса уникален, по-интерактивен и удобен за студентите. Анализът на натрупаните данни ни дава шанс да анализираме как учащите реагират на курсовете и да оценим ефективността на метода на преподаване, кои стратегии за обучение работят добре и кои не са полезни по отношение на постигането на целите на e-Learning [1].

Друга изследвана характеристика е как обучаващите предпочитат да достигат до търсената информация. Например, голяма част от учащите предпочитат видео уроците, оптимизирани за мобилни устройства. На база на натрупаните данни и техния анализ преподавателите могат да променят начина, по който доставят учебното съдържание.

С достъпа до интернет се променя и модела на оценяване. Студентите вече решават казуси и работят по проекти, базирани на изследователска дейност, вместо да отговарят на въпроси с множество възможности за избор. Чрез анализ на натрупаните данни, може да се проследи как се справят с предизвикателства през целия модул на обучение от началото до края. Анализът на данните, натрупвани в курсовете за електронно обучение дава възможност да се промени модела на изпитване и да се проектират модули, които да отговарят на индивидуалните нужди на обучаемия, относно търсене на необходимата информация [4].

Базирайки се на натрупани данни от работата на система за електронно обучение с различни потребители, прилагайки средства от областта на науката за данните, могат да се взимат различни решения относно обучението [13]. Отворени проблеми в направлението са свързани с оптимизиране на достъпа до данните, чрез прилагане на агенти за извличане на знания и нови техники за анализ на данни.

Интересът на обучаемите, активното им участие в процеса на усвояването на знания и придобиването на умения, в голяма степен могат да бъдат повлияни от качеството на използваната обучаваща среда.

Заключение

Важно условие за успешното развитие на икономиката на настоящия етап е способността да се улавят и анализират големи масиви и информационни потоци. Налага се мнението, че овладяването на ефективни методи за анализ на големи данни е условие за индустриална революция. Четвъртата индустриална революция [14] е тясно свързана със следващото поколение Интернет, което трябва да гарантира, че огромният потенциал на изкуствения интелект, виртуалната реалност, връзката с физическия свят, машинното обучение, повсеместните мрежи от хора и машини се използват пълноценно за всеобщо подобряване качеството на живот и допринасят за изграждането на устойчиви общества.

Литература

[1] Agarwal H., G.N.Pandey, „Impact of E-Learning in Education", International Journal of Science and Research (IJSR) ISSN: 2319-7064.
[2] Bernard Marr., „Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance". John Wiley & Sons Ltd, 2015.
[3] Mark Hornick, Erik Marcade, Sunil Venkayala, Java data mining: strategy, standard, and practice, A practical Guide for Architecture, Design, and Implementation, 2006.
[4] Orozova, D. Appropriate e-test system selection model, Comptes rendus de lAcademie bulgare des Sciences, Vol 72, No. 6, 811-820.
[5] Russell, S and P. Norvig. Artificial Intelligence: A Modern Approach, Prentice Hall, Upper Saddle River, New Jersey, 1995.
[6] Venkatram K., Geetha Mary A., Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications, Cybernetics and Information Technologies, Vol. 17(2), 2017, 3-27, ISSN: 1311-9702;
[7] Popchev I., D. Orozova, Towards Big Data Analytics in the E-learning Space, Cybernetics and Information Technologies, Vol. 19(3), 2019, 16-24, ISSN: 1311-9702;
[8] Gramatova K., S. Stoyanov, E. Doychev, V. Valkanov, Integration of eTesting in an IoT eLearning ecosystem: Virtual eLearning Space, BCI 2015, 14:1-14:8.
[9] Stancheva, N., A. Stoyanova-Doycheva, S. Stoyanov, I. Popchev, V. Ivanova. An Environment for Automatic Test Generation. – Cybernetics and information Technologies, Vol. 17, No. 2, Sofia, 2017, 183-196.
[10] Stancheva, N., A. Stoyanova-Doycheva, S. Stoyanov, I. Popchev, V. Ivanova. A Model for Generation of Test Questions. – Comptes rendus de l'Academie Bulgare des Sciences. Vol. 70, No. 5, 2017, 619-630.
[11] Stoyanov S., I. Ganchev, I. Popchev, I. Dimitrov (2010) Request globalization in an infostation network, C. R. Acad. Bulg. Sci., 63(6), 901-908.
[12] Stoyanov S., V. Valkanov, I. Popchev, A. Stoyanova-Doycheva, E. Doychev (2014) A Model of context – aware agent architecture, C. R. Acad. Bulg. Sci., 67(4), 487–496.
[13] Орозова, Д., С. Стоянов и И. Попчев, „Виртуално образователно пространство" в Научна конференция с международно участие „Знанието – източник на иновация", БСУ, Бургас, 2013, pp. 153-159, ISBN 978-954-9370-99-7.
[14] Шваб, К. Четвъртата индустриална революция (превод от английски). Издателство „ХЕРМЕС". Пловдив, 2016, 240, ISBN 978-954-26-1630-6.
[15] https://orange.biolab.si. [Online]. https://orange.biolab.si/training/introduction-to-data-mining/